

A Brief Overview of Explanations in Expert and Recommender Systems

Cilliers Pretorius
prtpie003@myuct.ac.za
University of Cape Town
Rondebosch, South Africa

ABSTRACT

Explanations are a challenging aspect of expert and recommender systems. With analysis of the types of explanation knowledge and user habits, insight is gained in designing an explanation tool for likely users. Defining the quality of explanations according to aspects such as transparency and persuasiveness, it is clear that explanations and human-computer interaction is tightly intertwined. While symbolic expert systems have thoroughly researched explanation tools, non-symbolic systems such as neural networks prove otherwise and have a lot of research potential. Similarly, recommender systems will benefit from further research into explanations. There is clear motivation for including explanation tools in expert and recommender systems.

CCS CONCEPTS

- **Information systems** → *Expert systems; Recommender systems;*
- **Computing methodologies** → *Knowledge representation and reasoning;*

KEYWORDS

Explanation, Expert system tools and techniques, Recommender systems, Artificial Intelligence

1 INTRODUCTION

As computational devices grow in power and decrease in cost, more uses are finding their way into everyday use in all areas. One particular area of importance is in providing explanations to support and justify a decision made by a system.

Explanations in expert systems (computer applications meant to provide the user with a decision or recommendation based on a set of rules and facts used to deduce new facts with user input) were first explored in the mid-1970s by Stanford's MYCIN medical diagnosis expert system. [12] For a while, MYCIN was unique in providing explanations to the users, with few developers and users opting to include explanation tools in their expert systems.

However, a large body of research was produced afterwards, [8] chief among that led to a significant increase in understanding of explanation as it pertains to computer-based systems. In particular, the epistemological model of knowledge put forward by [8] is of importance.

Soon, this explanation and justification was considered to be a key part of an expert system and the acceptance of an expert system's recommendations.[33] This, in part, deals with the skepticism that was inherent in users of early expert systems, who had no explanation components at times.[4]

There is a wide array of applications in artificial intelligence (AI) beyond expert systems that benefit from explanations, an ubiquitous example being the recommender systems that pervade the Internet.[16] Given the general black box nature of these systems, explanations allow for greater user acceptance of recommendations and thus further accomplish the goals of these recommender systems.[16]

In particular, neural networks, which are often used in recommender systems ([10], [7]), are another example of systems that are very opaque to users, and will greatly benefit from explanations, again achieving greater success in the acceptance of the recommendations.

2 EXPLANATION DEFINED

The concept of explanation is very easy to understand intuitively, but a formal definition is difficult to produce. Different authors in different fields of study all use distinct definitions, and no clear consensus exists between these authors. [6] These differing definitions and uses are thoroughly explored by [6]. For the purposes of this paper, we will use the definition as described by [21]: Explanations are a description of the causal mechanisms that result in the outcome, where the outcome is the recommendation provided by the expert or recommender system. In essence, explanation is the answer to the question, "How or why does the system provide a particular result?"

2.1 Types of explanation

[8] defined an epistemology for explanation that centers around three types of knowledge. These types of knowledge generally correspond to the questions a user is likely to ask when using an explanation tool for a expert or recommender system. These types of knowledge and their definitions will be visualised in terms of a spellchecker program that has an explanation component – reasonably common with both Microsoft Word and Grammarly.com both having such components.

2.1.1 Terminological knowledge. Terminological knowledge is the knowledge of the domain, its concepts and the relationships that form it. A user must understand the terminology of the domain to understand the domain. In the example of the spellchecker, an example of this is knowledge that results from a user asking, "What does it mean that the tense is wrong?" In essence, this is knowledge that is required to understand the explanation provided by the knowledge system.

2.1.2 Justification knowledge. Justification knowledge is knowledge of a particular system's reasoning. It is knowledge of facts that can be deduced from a given set of facts and rules, what entailment

resulted in a particular outcome. In the spellchecker example, this knowledge is the answer to a question such as, "Why is the use of this pronoun wrong?" This type of knowledge is the most commonly used type of explanation when using explanation tools.[2]

2.1.3 Trace knowledge. Trace knowledge is knowledge of how a particular system operates. It is the algorithms and particular steps the system followed to reach a specific conclusion or recommendation. In the running example, this is the answer to a question such as, "How is it decided that this verb is used incorrectly?" This type of knowledge is vital when testing the system and understanding how an incorrect conclusion or recommendation is reached.

2.1.4 Strategic knowledge. Strategic knowledge is knowledge about how the system uses or requires some particular bit of information. [37] In the spellchecker example, a suitable question would be, "Why is it necessary to say what variant of English I use?" Strategic knowledge is sometimes considered a subsection of justification knowledge.

2.2 Users of explanation tools

There is some research on how the types of explanation is used differently by novices and experts, as well as determining the reasons the users give for using the types of explanation. Generally, the two main use categories for explanations are learning, and problem-solving.

A key distinction that occurs in explanation usage is between expert users and novice users. This is merely in the manner in which explanations are accessed and not the volume or frequency thereof. Indeed, a study by [37] has shown that novices and experts access about the same amount of explanations.

Novice users are likely to use justification knowledge more than any other type of explanation.[11] Trace and strategic knowledge is rarely used by novices. In particular, novices tend to use explanations more when they have the goal of learning.[15] Similarly, when a user is ignorant about some aspect needed to understand and add to the problem-solving ability, the user is likely to request explanations.

Expert users tend to use explanations mainly when they disagree with the system's recommendations. [11] Expert users are unlikely to use explanation tools where the difficulty to access the explanation exceeds the perceived value of the explanation. [22] This, according to [22], leads to users being unaware of the value that the explanation systems add to the expert system compared to automatic explanations. Automatic explanations are explanations are embedded within the usual human-computer interaction and does not have to be invoked by the user in some manner. [15]

2.3 Evaluating explanations

When defining explanation, one must also take care to define a method of evaluating an explanation and its acceptability. Given that explanations are aimed at the user, this requires the human factor to be taken into account. [34] provides an interesting generalised argument from the psychological and philosophical viewpoints on the *theory of explanatory coherence* that "provides a psychologically plausible account of how people evaluate competing explanations."

With this paper's more specific focus, [35]'s list of criteria for evaluating explanations in recommender systems will instead be described. Much of these criteria can be traced back to human-computer interaction fundamentals as set out by Nielsen.

2.3.1 Transparency. Explanations contribute to transparency if they explain how a system works or why a specific result is returned. Transparency is desired because it allows for "Visibility of System Status" as defined by [23].

2.3.2 Scrutability. Scrutability is where explanations allow the user to query and understand why a particular result is returned. In recommender systems, scrutability allows the user to correct the assumptions made that resulted in incorrect recommendations. This allows for the heuristic of User Control. [23]

2.3.3 Trust. An explanatory system is considered trusted if its competence can be relied on. A trusted explanation is more likely to be accepted by the user. It must be noted that trust is difficult to measure with a wide variety of factors, the interface design in particular, influencing the perceived trust of an explanation.

2.3.4 Effectiveness. An effective explanation is an explanation that helps users make *better* decisions. This is highly dependent on the accuracy of the system that produces the explanation, but effective systems allow for users to be satisfied without more in-depth explorations of the reasoning and domain knowledge.

2.3.5 Persuasiveness. Persuasiveness is a measure of how likely users are to adhere to a system's recommendation after the explanation. A more persuasive explanation is generally considered to be a better explanation, although there are exceptions.

2.3.6 Efficiency. The efficiency of an explanation is simply how long it takes for users to make their decisions based on the explanation. Alternately, it can be seen as how many explanations are needed for the user to be satisfied. The more efficient an explanation, the better that explanation is.

2.3.7 Satisfaction. The users' satisfaction with an explanation system is often difficult to measure independently of the rest of the system. However, a satisfactory explanation leads to repeat use of the system and greater acceptance of the recommendations.

3 EXPLANATION TECHNIQUES

The techniques used to extract an explanation from a system differ depending on the type of system used and its specific implementation.

3.1 Expert systems

Most expert systems utilise symbolic techniques, i.e., they use explicit symbols stored and the logical manipulation of these produce the output.

MYCIN is the prototypical example of a *rule-based system*. With a set of rules, a justification explanation is easily generated by just producing a log of the rules that were invoked in a particular execution. [32] This is considered a form of feedback.

MYCIN also allows users to ask for strategic explanations [12], such as why a particular question needs to be answered. Using a

feedforward mechanism, the system traces the followup rules to determine what the question allows for.

These rule traces are helpful when developing and debugging a system, since it allows for easy visualisation of both the system code and the knowledge base. The drawback of rule-based systems is that justification knowledge needs to be explicitly included with the set of rules. If the existence of a rule is not explained explicitly, a rule trace might not be able to explain why a particular rule exists and has given rise to a specific result. [8]

A second form of symbolic system is *object-based systems*. These systems use knowledge stored in a hierarchical collection of objects. Objects can have data attached to them and can inherit data from other objects. This hierarchy allows for relationships between objects and knowledge to be easily identified. A good example of this type of system is Protégé. [17]

In contrast to rule-based systems, object-based systems aren't limited in what types of knowledge they can provide in explanation. [1] Since objects have knowledge attached to them, terminological knowledge is straightforward to provide in explanation. Similarly, with relationships also being an object [17], it also contains data and therefore justification knowledge and strategic knowledge can be provided separate from the domain knowledge of the system. Trace knowledge is easily explained by using a log of the system executing. This will supply what objects and relationships were used in order to produce a result.

An alternative approach to expert systems uses artificial neural networks (ANNs) and other forms of AI instead of the rigid, symbolic rules mentioned above. This non-symbolic approach has found great appeal in especially pattern recognition such as automatic image or speech recognition. ([24], [27])

ANNs operate by applying a system of weights to all input required or available for the task. This system of weights is calibrated by many training examples. Given that this system of weights is essentially a matrix of numbers, it is not easy to extract explanations from this type of system. Therefore, despite the capabilities provided by ANNs, the difficulty in explaining why the ANN produced a result has prevented its use in many cases. [3]

There have been efforts to extract explanations from ANNs, but these have not always been able to provide the types of explanations as discussed in section 2.1.[29] provides a way to determine what a neural network does, as well as how the structure allows for the function. In other words, neural networks can be explained with some trace and strategic knowledge types.

[36] produced a method to extract symbolic rules from trained neural networks, thereby allowing the same explanation techniques used for symbolic systems to be used on neural networks. This method does mean terminological and justification knowledge types can be difficult to explain unless explicitly stated, similar to how the justification knowledge needs to be explicitly included when using rule-based systems.

3.2 Recommender systems

Recommender systems are systems designed to advertise items to consumers that they are likely to enjoy, whether those items are objects on sale or videos on YouTube. These systems are aimed at helping people to make decisions on what to do next. ([25], [26])

The two most common approaches to recommender systems are content-based, and collaborative filtering (CF). There are other methods, and most modern applications utilise a hybrid approach between two or more of the above methods [26]. For the purposes of this paper, only content-based and collaborative filtering will be discussed.

3.2.1 Content-based. A logical method of suggesting new content for a user would be to look at the previous content the user has liked. In other words, suggest items similar to the items the user has bought in the past. This is the essence of content-based recommender systems.[19]

An end user of these systems will be interested in the justification and, less so, the terminological knowledge explanations for the item. Explanations would be relatively straightforward to generate – "Because the user showed interest earlier in similar items, this item is suggested" in the case of justification knowledge, while the terminological knowledge is inherent in the item as in object-based symbolic systems.

However, for a service such as YouTube, the history for a user can be extremely large and might take prohibitively long to generate an explanation. Also, in conjunction with other methods of recommendation, an explanation is considerably less easy to generate.

3.2.2 Collaborative filtering. An alternate approach to recommender systems is to collect data on the users, then suggest items to a user that was enjoyed by users similar to them. This similarity is determined by the user's profiles, and often their history is included to make for a hybrid approach.

Again, an end user will be interested in the justification and terminological knowledge type of explanations. The terminological data is inherent to the items, but the justification knowledge can produce problems. As [16] notes, the recommendations in CF are often based on incomplete data. This incompleteness results in inaccurate recommendations.

As mentioned in section 2.3.2, explanations can allow for scrutability and therefore inform the user why the system came up with the recommendation. It can also allow the user to add or adjust their information to generate a more accurate recommendation.[35] However, with recent privacy issues and regulations regarding personal data, it can be difficult to provide clear justification knowledge to a user. Since it relies on other users' personal data, it is not available to other users, and hence explanations can be difficult to provide.

4 MOTIVATION FOR EXPLANATIONS

With the relative difficulty of creating adequate explanation facilities, some might question why bother in the first place? In the 1980s, expert systems saw a rapid proliferation to the extent that within a decade, more than 300 of the *Fortune 500* companies were using at least one expert system. [13]

Explanations are now considered a key component of most knowledge-based systems used by professionals. [18] Generally, the greatest benefit of explanations is the increased user acceptance of the system's recommendations, although this is not always the case. According to [28], including explanations in a system leads to a considerable increase in user acceptance of the system and

its recommendations, while [14] didn't find a difference in user acceptance or performance when explanations were included.

As a more definitive argument for the benefits of explanations, when novice users use explanations, they solve problems faster and more accurately than when not using explanations.[14] When expert users have explanations available, they accept the system's recommendations more readily and more fully. [2]

Explanations lead to greater trust in the system, and with this greater trust comes an increased likelihood that the system will be used repeatedly and that its recommendations will be followed. [32] Improved belief leads to solving problems using the system more accurately and in a shorter time. ([2], [15], [37])

Recommender systems is another aspect of computing that benefits from explanations. Explanations – especially trace and strategic explanations – improve the transparency of the system. [5] Improved transparency is considered important to users of recommender systems.[30]

Another aspect to consider in recommender systems is the correctness of its recommendations. Unlike most expert systems, there is generally little active input from the user in these recommender systems and therefore inaccuracies can occur. Having explanations available allows the user to understand how the system works and allows for a natural segue to correcting the inaccuracies. [31]

As with expert systems, explanations in recommender systems allow for the users to adhere more closely to the suggestions. In recommender systems, this translates into users more often buying the recommended item. These explanations can convince users into buying items they would not have bought otherwise. ([9], [16]) It is interesting to note that unlike expert systems, the ideal 'persuasiveness' of a recommender system is not to be as successful as possible. If a recommender system is too successful in the short term, it is possible that the user will stop using it after buying too many items they weren't interested in. Thus, too much short-term success could lead to no long-term success.

Explanations in recommender systems benefit the system by allowing the user to be more efficient. These explanations can allow the user to quicker decide between similar products and understand the association between them. [20]

It is clear how explanations benefit recommender systems and increased acceptability of the recommendations results in increased profitability of the recommender system. Similarly, greater acceptance of expert system recommendations can result in more accurate decisions being made faster.

5 CONCLUSIONS

This paper explored explanation and its facets as it pertains to expert and recommender systems. A definition of explanation was provided, and the various types of explanation were analysed and explained using the example of a spellchecker. How users of differing skill levels use explanations is of interest and offers insight into how to design an explanation tool depending on the expected users.

An overview of how to evaluate an explanation was given. From this, it can be seen that the value of an explanation is quite heavily intertwined with human-computer interaction and its interface. A poor interface is likely to ruin even the best explanation tools.

A broad overview of expert systems' explanations was explored. Explanations for symbolic systems are a mature area of research and is well-defined. It proves to be useful for both novices and experts, and will remain relevant for the foreseeable future.

In contrast, non-symbolic systems such as neural networks have very little scope for easy explanations. At the same time, with the recent proliferation of artificial neural networks and their pattern-recognition abilities, explanations will be invaluable in deciphering the black-box model that ANNs consist of. There has been some research on this topic, but it will remain a fruitful topic for a long time yet.

Recommender systems have become ubiquitous on the Internet. An early approach in the form of content-based recommendations proved easy to provide explanations for. However, the modern approach is to combine several methods into a hybrid approach. This proves challenging to provide explanations for. With its ubiquity, more research in this area would be welcome.

There is clear motivation for explanations in both expert and recommender systems. Explanations improve the user adherence to the system suggestions and therefore improve the system performance. There has been thorough research on this topic, by several researchers.

Explanations are a key part to keep in mind when designing expert or recommender systems, with the potential to greatly increase the success of the system.

REFERENCES

- [1] Janice S Aikins, John C Kunz, Edward H Shortliffe, and Robert J Fallat. 1983. PUFF: an expert system for interpretation of pulmonary function data. *Computers and biomedical research* 16, 3 (1983), 199–208.
- [2] Vicky Arnold, Nicole Clark, Philip A Collier, Stewart A Leech, and Steve G Sutton. 2006. The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. *Mis Quarterly* (2006), 79–97.
- [3] Bart Baesens, Rudy Setiono, Christophe Mues, and Jan Vanthienen. 2003. Using neural network rule extraction and decision tables for credit-risk evaluation. *Management science* 49, 3 (2003), 312–329.
- [4] Dianne C Berry and Donald E Broadbent. 1987. Explanation and verbalization in a computer-assisted search task. *The Quarterly Journal of Experimental Psychology* 39, 4 (1987), 585–609.
- [5] Bruce Buchanan. 1984. Rule based expert systems. *The MYCIN Experiments of the Stanford Heuristic Programming Project* (1984).
- [6] Thomas K Burch. 1999. Computer modelling of theory: explanation for the 21st century. *PSC Discussion Papers Series* 13, 4 (1999), 1.
- [7] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishu Aradhya, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. ACM, 7–10.
- [8] William J Clancey. 1983. The epistemology of a rule-based expert system—a framework for explanation. *Artificial intelligence* 20, 3 (1983), 215–251.
- [9] Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. 2003. Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 585–592.
- [10] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. ACM, 191–198.
- [11] Jasbir S Dhaliwal. 1993. *An experimental investigation of the use of explanations provided by knowledge-based systems*. Ph.D. Dissertation. University of British Columbia.
- [12] Richard O Duda and Edward H Shortliffe. 1983. Expert systems research. *Science* 220, 4594 (1983), 261–268.
- [13] John Durkin. 1996. Expert systems: a view of the field. *IEEE Intelligent Systems* 2 (1996), 56–63.
- [14] Martha M Eining and Patrick B Dorr. 1991. The impact of expert system usage on experiential learning in an auditing setting. *Journal of Information Systems* 5, 1 (1991), 1–16.

- [15] Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly* (1999), 497–530.
- [16] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.
- [17] Matthew Horridge, Holger Knublauch, Alan Rector, Robert Stevens, and Chris Wroe. 2004. A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools Edition 1.0. *University of Manchester* (2004).
- [18] Izak Benbasat Ji-Ye Mao. 2000. The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *Journal of Management Information Systems* 17, 2 (2000), 153–179.
- [19] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*. Springer, 73–105.
- [20] David McSherry. 2005. Explanation in recommender systems. *Artificial Intelligence Review* 24, 2 (2005), 179–197.
- [21] Eugene J Meehan. 1968. *Explanation in social science: A system paradigm*. Dorsey Press.
- [22] Kathleen Ellen Moffitt. 1989. An empirical test of expert system explanation facility effects on incidental learning and decision-making. (1989).
- [23] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 249–256.
- [24] Yohhan Pao. 1989. Adaptive pattern recognition and neural networks. (1989).
- [25] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Commun. ACM* 40, 3 (1997), 56–59.
- [26] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer, 1–35.
- [27] Fred Richardson, Douglas Reynolds, and Najim Dehak. 2015. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters* 22, 10 (2015), 1671–1675.
- [28] Thomas Roth-Berghofer and Bjorn Forcher. 2011. Improving understandability of semantic search explanations. *Int. J. Knowledge Engineering and Data Mining* 1, 3 (2011), 216–234.
- [29] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems* 28, 11 (2017), 2660–2673.
- [30] Rashmi Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems*. ACM, 830–831.
- [31] Frode Sørmo, Jörg Cassens, and Agnar Aamodt. 2005. Explanation in case-based reasoning—perspectives and goals. *Artificial Intelligence Review* 24, 2 (2005), 109–143.
- [32] William R Swartout. 1983. XPLAIN: A system for creating and explaining expert consulting programs. *Artificial intelligence* 21, 3 (1983), 285–325.
- [33] Randy L Teach and Edward H Shortliffe. 1981. An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research* 14, 6 (1981), 542–558.
- [34] Paul Thagard. 2006. Evaluating explanations in law, science, and everyday life. *Current Directions in Psychological Science* 15, 3 (2006), 141–145.
- [35] Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*. Springer, 479–510.
- [36] Geoffrey G Towell and Jude W Shavlik. 1993. Extracting refined rules from knowledge-based neural networks. *Machine learning* 13, 1 (1993), 71–101.
- [37] L Richard Ye and Paul E Johnson. 1995. The impact of explanation facilities on user acceptance of expert systems advice. *Mis Quarterly* (1995), 157–172.